This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAES.2018.2879533, IEEE Transactions on Aerospace and Electronic Systems

Sequential Sensor Selection and Access Decision for Spectrum Sharing

Jihyun Lee, *Student Member, IEEE* and Eylem Ekici, *Fellow, IEEE*

***Abstract*—We develop an algorithm for sequential sensor selection and channel access decision in CR (cognitive radio) networks under shadowing with the objective of maximizing total number of correct decisions on channel availability over finite time periods. The problem is formulated as a MAB (Multi-Armed Bandit) problem and our index-based heuristic algorithms is shown to achieve sub-linear accumulated regret in time period T. We also evaluate the performance of the proposed algorithms with numerical examples.**

***Index Terms*—Cognitive radio, Sensor selection algorithms, channel access decision, Multi-armed bandit problems**

I. Introduction

To meet the exponentially growing demand in mobile data, spectrum sharing between wireless communication systems and radars has become an emerging area of research [1], [2]. Radar bands are especially attractive for spectrum sharing because of its large total bandwidth (1GHz below 10GHz) and radar's sparse geographical use. In fact, the US designated the 3.55 GHz radar band to be a shared band and established the Citizens Broadband Radio Service (CBRS), allowing commercial systems to use the band previously used mainly by radar systems (e.g., SPY-1 and SPN-43). In the 5 GHz UNII bands (i.e., U-NII-2B (5350-5470 MHz) and U-NII-4 (5850-5925 MHz)), wireless applications are also allowed to coexist with radars [2]. However, the regulation is very conservative requiring a high sensitivity level of secondary users to avoid causing interference to radars, which makes it hard for the band to be efficiently utilized by other wireless applications.

Therefore, to protect licensed users (primary users, PU) while achieving high spectrum utilization by the unlicensed users (secondary users, SU), it is important to improve the sensing performance of the SU network. The sensing performance can be improved by increasing the number of sensors and/or the sensing time [4]. However, it is not desirable to have all sensors perform spectrum sensing every time due to the energy constraint on the handheld devices. Moreover, for a centralized SU network, sensing results should be sent back to the controller, which results in high overhead in the control channel. Therefore, it is more desirable that a subset of sensors are selected for sensing at a time. Similarly, the sensing time is also limited because SUs should remain silent during sensing periods, which wastes precious communication opportunities. To maximize the sensing performance under these limitations, optimal solutions of the sensor selection and channel access decision have been extensively studied [3]–[7]. When statistical information such as the distribution of received PU signal at SU side is known, it is well known that the optimal channel access decision rule can be obtained by using the likelihood ratio test (LRT) of binary hypothesis testing [8] for a given sensor selection. Finding an optimal sensor selection generally includes applications of LRT to all possible sensor candidates.

However, the presence of shadow fading poses a challenge in designing an optimal sensor selection and access decision rule. Depending on the location, each SU experiences different channel conditions from the PU transmitter. Even though the path-loss effect on an SU network is assumed to be known (or can be computed from location information), shadow fading effects still remain. Therefore, channels between the PU transmitter and SU receivers are not known to the SU network controller a priori and can only be learned through sensing feedback [4]. The feedback information, however, is also limited as only a subset of SUs can be selected for sensing at a time and the sensing results are contaminated by noise. Moreover, only the sensing results measured during the active period of PU contain information on the channel quality between PU and SU.

To cope with such uncertainties resulting from lack of prior knowledge and limited feedback, we show that the problem can be considered as an instance of the multi armed bandit (MAB) model in which each SU is considered as an arm. Thus, the objective is to find a sensor selection and a channel access decision rule which maximizes sensing performance in terms of the total number of correct decisions on the channel availability over a finite time period. In contrast to other works where the statistics of PU channel process is unknown while the sensing is error-free, in our problem, the error probability of each SU is initially unknown and the statistics of PU activity is assumed to be known, e.g., i.i.d. Bernoulli process with a known mean. This differentiates our problem from earlier works in the following aspects. Firstly, ***the stochastic processes in our problem have a hierarchical structure*** where the distribution from which the sensing samples are taken at each round depends not only on the sensor selection but also on the state of the PU (idle or busy). So, even for a given sensor selection, at each time t, the sensing samples are taken from a mixture of two Gaussian distributions, which does not fall into the standard reward distributions such as the exponential family [9], [26] or an independent Markov machine [10], [14]–[18]. An EM (Expectation Maximization) algorithm for Gaussian mixture models is used for parameter learning in [23], but the sensor selection problem is not

This work has been supported by the NSF under Grants CNS-1421576 and CNS-1731698.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAES.2018.2879533, IEEE Transactions on Aerospace and Electronic Systems

considered. Secondly, unlike other MAB formulations which issue a reward depending on the arm selection, in our problem, the reward depends both on the sensor selection and the channel access decision. Therefore, even if the sensing samples are taken from an i.i.d. distribution for a given selection independent from the history, ***the reward process in our problem can be no longer i.i.d. due to the dependency of the reward on the access decision rule***. We will show how we benefit from the i.i.d property of the sampling space to tackle this problem later in Section IV. For a detailed discussion of the non-i.i.d. rewards, see Section IV.

Our contributions are three fold.

- We formulate the sequential sensor selection and channel access decision problem in the context of MAB problems to leverage an exploitation-exploration trade-off, which includes uncertainties on the sensing abilities of SUs due to unknown shadow fading. The received PU signal on the SU side is modelled to have a Gaussian distribution and the unknown shadowing effect is incorporated as an unknown mean of the distribution.
- We propose an upper confidence bound (UCB) based sensor selection algorithm and a sample mean based channel access rule. It is shown that the number of selections of sub-optimal SUs increases at most ***logarithmically*** in time horizon T, i.e. $O(\log T)$, which is known to be order optimal. Combined with our channel access rule, we finally show that our algorithm has a ***sub-linear*** regret growth rate, which is at most $O(T^{\frac{2}{3}+\epsilon})$.
- The performance of our algorithms are evaluated through simulations under different settings of unknown parameters. Also, we compare the performance of our sensor selection algorithm with UCB-1 [11], an alternative solution approach to the sensor selection part of our problem. For the comparison of selection algorithms, fixed threshold test is used for the channel access decision.

A. Related Work

1) Sensor selection and channel access decision: Theoretically, finding an optimal channel access decision rule for spectrum sharing is the same as a binary detection problem widely studied in sensor networks whose optimal solution is obtained by LRT. However, the LRT-based solution is applicable for the cases where the measurement model (the probability distribution of the sensing measurement under each hypothesis) is known to the decision maker and it becomes computationally intractable as the number of sensors increases. In [6], the measurement distribution of the active PU channel is modeled as a correlated Gaussian under the assumption of the log-normal random shadowing with all parameters known. They provide an approximation decision rule based on deflection criteria to tackle the computation issue of LRT. On the other hand, in [4], the authors show that the optimal rule can be obtained from LDA (Linear discriminant analysis) if the shadowing is considered as a fixed parameter (which is a specific realization of a random variable) as in this case, the measurement distributions have a common covariance matrix. An optimal sensor selection can be found by repeatedly applying LRT (or its approximation methods) to each candidate for a known measurement model. In the cases where LDA assumptions are satisfied, the sensor selection problem reduces to finding a sensor set which have the best PU transmitter-to-SU channel quality in the mean [4]. In [5], a Bayesian decision rule-based access decision algorithm is proposed to maximize system throughput when the probability of mis-detection and the probability of false alarm of each SU is known. In this paper, in contrast to [4]–[6], we consider learning the optimal sensor selection and access decision when neither the measurement model nor local sensing ability such as the probability of mis-detection or false alarm is initially unknown.

2) Multi-armed bandit: MAB models have been widely adopted for a large class of sequential optimization problems [9]–[22], providing learning solutions during operation time. In classical MAB problems [9], [11], when a player pulls an arm, it offers an i.i.d. random reward drawn from its associated distribution with an unknown mean. At each time, the decision maker chooses an arm, aiming to maximize the total expected reward in the long run. This problem involves the well-known tradeoff between exploitation (maximizing instantaneous reward) and exploration (learning unknown reward statistics). MAB usually focuses on a quantity termed cumulative expected regret. The cumulative expected regret of a sequence of decisions is simply the cumulative difference between the expected reward of the options chosen and the maximum reward possible. In [11] and [26] it is shown that the confidence bound-based algorithms achieve logarithmic regret uniformly in time for bounded reward distributions and Gaussian reward distributions. There are also many variants of the classical MAB models including Markov rewards and multi-player multi-arm scenarios. For example, an application of restless MAB formulation to wireless downlink scheduling problem with ARQ-based CSI feedback can be found in [18]. In [12], [13], multi-band spectrum sensing and access problem in CR is formulated as a MAB problem where each PU channel is modeled as an arm with unknown i.i.d. Bernoulli reward process. Extensions to Markov PU processes and decentralized multi-user multi-band sensing scenarios can be found in [10], [14]–[17]. However, most of the CR literature focuses on learning unknown channel availability statistics, assuming perfect sensing ability, i.e. an SU can acquire an error-free knowledge on the state of the sensed channel. In [24], [25], the authors investigated the impact of sensing errors on the channel selection algorithms, but optimization over sensing abilities is not considered.

The rest of the paper is organized as follows. In Section II, we formulate the problem and present the system model. In Section III, we design a sensor selection algorithm assuming that a genie-aided LRT decision rule is employed for each selection. In Section IV, we provide a joint design of sensor selection and channel access decision. Finally, we evaluate the performance of our proposed algorithms in Section V and conclude the paper in SectionVI.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAES.2018.2879533, IEEE Transactions on Aerospace and Electronic Systems

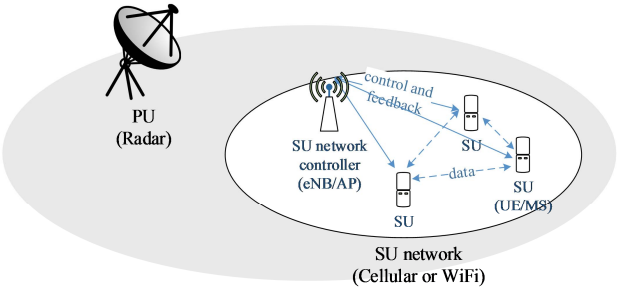


Fig. 1: Spectrum sharing scenario

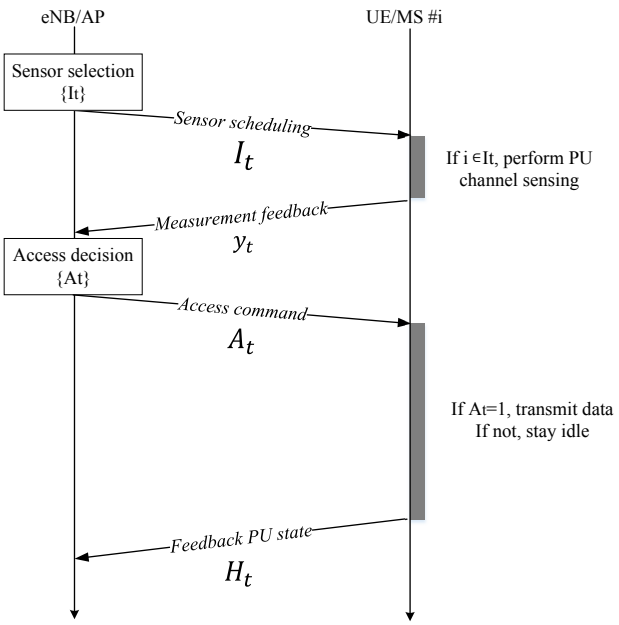


Fig. 2: SU network protocol for spectrum sharing

Sensor selection	UCB based sensor selection	
Access decision	Fixed threshold test (Genie-aided)	Threshold test based on sample mean
	Alg. 1. UCB-FT (Sec.III)	**Alg. 2.** UCB-LLR (Sec. IV)

Fig. 3: Summary of the algorithms

II. System model and Problem formulation

A. Network model

As shown in Fig.1, we consider a centralized SU network with N SUs indexed by $i \in \{1,2,\cdots,N\}$. Each SU is assumed to be equipped with a sensor and SUs and sensors are used interchangeably in this paper. At the beginning of each time slot indexed by $t \in \{1,2,\cdots,T\}$, the SU network controller schedules an SU to sense the channel (sensor selection). On receipt of the schedule, the selected SU performs sensing on the PU channel and reports its local measurement to the network controller. We assume an SU uses energy detection for spectrum sensing because the PU signal features such as pilots and preambles are unknown to the SUs [4]. After the network controller collects local sensing data, it make a decision on channel availability (channel access decision). If the controller's decision on channel access is positive, SUs access the channel for the remaining period of the time slot. At the end of the time slot t, we assume that the controller obtains the information of true channel state at time t. The SU network controller does not perform sensing and control commands and feedbacks are transmitted using a separate (dedicated) control channel for a reliable network management. The procedure and the algorithms to be used are summarized in Fig. 2 and Fig. 3, respectively.

Let us denote by $\mathbf{y}_t$ the distribution of signal strength of PU measured by SUs at time t. Under the same assumption as in [3], [4], [7], $\mathbf{y}_t$ has a multi-variate Gaussian distribution:

$$\mathbf{y}_t \sim \begin{cases} \mathcal{N}(0,\sigma^2 I) & \text{when } H_t = 0 \\ \mathcal{N}(\boldsymbol{\mu},\Sigma) & \text{when } H_t = 1 \end{cases} \quad (1)$$

where H_t represents true channel state at time t and when it is 0 (1), it means the channel is idle (busy). $\boldsymbol{\mu}$ is an $N \times 1$ vector, i-th component of which represents received signal strength at SU i which is determined by the PU transmit power and the channel gain between the PU transmitter and the SU i. To incorporate shadowing effect into the model, in [3], [7], it is assumed that under H_1, all SUs have the same mean μ and the variance is the summation of the variance of the noise and the shadowing. (Note in this case the Covariance matrix of y^t under $H_t = 0$ and $H_t = 1$ is not the same). In this paper, we assume that $\boldsymbol{\mu}$ is unknown. This is a reasonable assumption because at a fixed location, shadowing effect is time-invariant and can be regarded as an specific yet unknown realization of a random variable (See [4] for details). Now the only randomness in the model comes from noise effect, we can assume the covariance matrix under H_1, Σ, is also equivalent to $\sigma^2 I$, where σ^2 is the known sampling noise variance. We note that the noise effect is independent across SUs. Also, we can assume the same noise variance for all SU because when the noise variance is different, we can define a new normalised random variable $z_i^t = \frac{y_i^t}{\sigma_i}$ and use z_i^t instead of y_i^t.

Using an appropriate shadowing model, we can also make use of our prior information on the unknown parameter $\boldsymbol{\mu}$. For example, under log-normal shadowing with a dB spread σ_0, we have a Gaussian prior distribution for μ in dB:

$$\mu \sim \mathcal{N}(\mu_0,\Sigma_0), \quad (2)$$

μ_0 can be computed from the path-loss model without shadowing. The randomness due to shadowing can be modeled in covariance by setting the diagonal component of Σ_0 to the shadowing variance σ_0^2 and the (i,j) off-diagonal components to $\rho_{ij}\sigma_0^2$, where ρ_{ij} is correlation coefficient between SUs i and j.

We assume that the process of PU channel availability is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAES.2018.2879533, IEEE Transactions on Aerospace and Electronic Systems

a Bernoulli process with known mean $\theta \in (0,1)$. When SU i is scheduled at time t, the sensing sample will be taken from one of the following distributions depending on the state of PU channel.

$$y_i^t \sim \begin{cases} \mathcal{N}(0,\sigma^2) & \text{when } H_t = 0 \\ \mathcal{N}(\mu_i,\sigma^2) & \text{when } H_t = 1 \end{cases} \tag{3}$$

μ_i is the i-th element of unknown $\boldsymbol{\mu}$ in (1). Note that the distribution of y_t given that SU i is selected at time t is equivalent to y_i^t, and the measurement model of (1) and (3) are independent of t under time-invariant shadowing. Since we assume σ to be known, the distribution of y_i^t is known under $H_t = 0$. The distribution of y_i^t under $H_t = 1$, however, is unknown with a single unknown parameter μ_i. Table I summarizes the notations used in the paper.

B. Reward

Let us denote by $\{I_t\}_{t=1}^T$ and $\{A_t\}_{t=1}^T$ a finite time sensor selection policy and an access decision policy, respectively. A policy I consists of a sequence of random variables $I_t \in \{1,2,\cdots,N\}$, indicating which SU has been selected for sensing at time t based on the history of observations $(I_1, y_1, A_1, H_1, \cdots, I_{t-1}, y_{t-1}, A_{t-1}, H_{t-1})$. A policy A consists of a sequence of binary random variables $\{A_t\}_{t=1}^T$, indicating whether to access the channel in time slot t or not, based on all the past informations including sensor selection I_t and its sensing result y_t. Let us define a reward collected at time t as follows:

$$X_t = \begin{cases} 1 & \text{if } A_t = H_t \\ 0 & \text{if } A_t \neq H_t \end{cases} \tag{4}$$

X_t is an indicator of whether the correct decision on the channel availability is made at the time t. Then,

$$\begin{aligned} \mathbb{E}[X_t] &= \mathbb{E}[\mathbb{1}(A_t = 1, H_t = 1) + \mathbb{1}(A_t = 0, H_t = 0)] \\ &= \mathbb{P}(H_t = 1)(1 - P_{MD}) + \mathbb{P}(H_t = 0)(1 - P_{FA}), \end{aligned} \tag{5}$$

where $\mathbb{1}(\cdot)$ is the indicator function and P_{MD} and P_{FA} is the mis-detection probability $\mathbb{P}(A_t = 0|H_t = 1)$ and the false alarm probability $\mathbb{P}(A_t = 1|H_t = 0)$ at time t, respectively. The expected reward can be interpreted as the expected system throughput (sum of the PU throughput and the SU throughput) at time t, if the capacity of PU and SU are the same and concurrent transmission of SU and PU is not allowed [5]. In fact, it is also possible to define the reward function as a weighted sum of the indicator functions of the possible combinations of A_t and H_t. For example, instead of (5), we can define $X_t = w\mathbb{1}(A_t = 1, H_t = 1) + (1-w)\mathbb{1}(A_t = 0, H_t = 0)$ for some $w \in (0,1)$ to balance the trade-off between the P_{MD} and the P_{FA}. The arguments in the following sections still hold, but for the exposition purpose, we will focus on (5). In the case of $w = 1$, the problem becomes a detection probability maximization, but we need to put a constraint on the maximum P_{FA} to avoid an undesirably conservative solution.

C. Objective

The objective is to design policies $\{I_t\}_{t=1}^T$ and $\{A_t\}_{t=1}^T$ for which the expected accumulated rewards collected up to time T, i.e. $\sum_{t=1}^{T} \mathbb{E}[X_t]$ is maximized. Equivalently, the expected regret up to time T, i.e. $R_{\text{opt}}^T - \sum_{t=1}^{T} \mathbb{E}[X_t]$ is minimized, where the optimal reward R_{opt}^T is the maximum expected reward gained by a genie to which all the parameters of the model is known. With the definition of reward X_t in (4), this can be interpreted as designing a sensing scheduling and an access policy which maximizes the expected number of correct decisions on channel availability over a finite time period T. We note that it is also possible that A_t is independent of past decisions and observations including sensor selection I_t and sensing result y_t. For example, at one extreme, we can design a deterministic access rule such that $A_t = 1$ for all t. However, the expected reward of this access policy will remain to be θ for all t. In the following sections, it will be shown how we can approach to the maximum expected reward by making access decisions based on the sensing results with a carefully designed sensor selection algorithm.

Symbol	Meaning
y_i^t	Received PU signal strength at i-th SU at time t
μ_i	Unknown mean of $y_i^t \|H_t = 1$
σ	Variance of $y_i^t \|H_t = 1$ and $y_i^t \|H_t = 0$
μ_i^t	Mean of prior distribution of μ_i
σ_i^t	Variance of prior distribution of μ_i
H_t	PU activity at time t
θ	Probability that PU is busy
I_t	SU selection at time t
A_t	Access decision at time t
η_i^t	Threshold used for access decision at time t when $I_t = i$
X_t	Reward collected at time t
R_{opt}^T	Optimal expected total reward up to time T with a known μ

TABLE I: Notation list

III. Sequential Sensor selection with fixed threshold test

A. Definition of Regret

In this section, we first consider a sensor selection policy $\{I_t\}_{t=1}^T$. It is assumed that the access policy uses a genie-aided fixed threshold test for each SU selection and the value of the threshold is given. Under this assumption, on receipt of y_t from SU I_t, the controller makes a channel access decision according to the following threshold test.

$$y_t \underset{A_t=0}{\overset{A_t=1}{\gtrless}} \eta_{I_t} \tag{6}$$

where $\boldsymbol{\eta}$ corresponds to the threshold of LRT, i.e. $\eta_i = \frac{\mu_i}{2} + \log\frac{1-\theta}{\theta}$. Then, the reward X_t given $I_i = i$ is $\mathrm{X}_i^t = \mathbb{1}(y_i^t \geq \eta_i, H_t = 1) + \mathbb{1}(y_i^t < \eta_i, H_t = 0)$ and successive scheduling of SU i will yield the rewards $X_i^1, X_i^2, \cdots$ according to some unknown i.i.d. distribution

$$X_i^t = \begin{cases} 1 & \text{with probability } \mathrm{p}_i \\ 0 & \text{with probability 1-p}_i \end{cases} \tag{7}$$

where $1 - p_i$ is the error probability of SU i. Using (3) and (6), p_i can be computed as follows:

$$p_i = (1-\theta)\Phi\left(\frac{\eta_i}{\sigma}\right) + \theta\Phi\left(\frac{\mu_i - \eta_i}{\sigma}\right) \tag{8}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAES.2018.2879533, IEEE Transactions on Aerospace and Electronic Systems

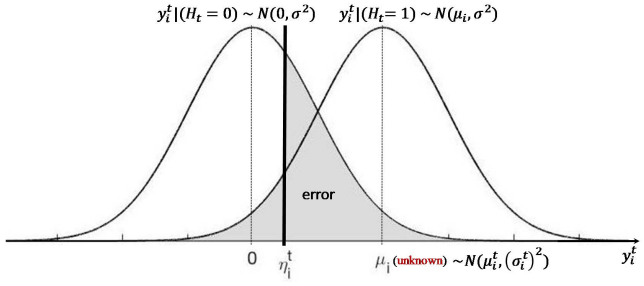


Fig. 4: Measurement distribution at SU i at time t: when $H_t = 1$, the sensing sample is taken from a Gaussian with unknown mean μ_i if $I_t = i$. μ_i has a Gaussian prior with mean μ_i^t and variance $(\sigma_i^t)^2$. Note that the instantaneous expected reward p_i at time t corresponds to the unshaded area of the plot which not only depends on μ_i (the selection) but also on the threshold η_i^t (access decision)

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal (see Fig. 4). Note that p_i is unknown as it is a function of unknown parameter μ_i. Now, we define the optimal reward R_{opt}^T. Let i^* denote the index of the SU with largest μ, i.e. $i^* = \text{argmax}_i\, \mu_i$. Let us call the selection of SU i^* the optimal selection. Then, the optimal sensor selection policy always selects SU i^*, yielding the total expected reward $R_{\text{opt}}^T = \sum_{t=1}^{T} \mathbb{E}[X_{i^*}^t] = p^*T$, where $p^* = (1-\theta)\Phi(\frac{\eta_{i^*}}{\sigma}) + \theta\Phi(\frac{\mu_{i^*}-\eta_{i^*}}{\sigma})$.

Therefore, the expected total regret at time T is defined as

$$\begin{aligned} R_{\text{opt}}^T - \sum_{t=1}^{T}\mathbb{E}[X_t] &= p^*T - \sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}[\mathbb{E}[X_t\,\mathbb{1}(I_t=i)|\mathcal{F}_{t-1}]] \\ &= \sum_{i=1}^{N} p^*\mathbb{E}[n_i^T] - \sum_{i=1}^{N}\mathbb{E}[X_i]\mathbb{E}[n_i^T] \\ &= \sum_{i=1}^{N}(p^*-p_i)\mathbb{E}[n_i^T] \end{aligned} \tag{9}$$

where, $\mathbb{1}(\cdot)$ is the indicator function, $n_i^T = \sum_{t=1}^{T}\mathbb{1}(I_t = i)$ and $\mathcal{F}_{t-1}$ is σ-field generated by previous decisions and observations. The second equality follows since $T = \sum_{i=1}^{N} n_i^T$, $\mathbb{1}(I_t = i) \in \mathcal{F}_{t-1}$ and X_i^t is independent of $\mathcal{F}_{t-1}$ (i.i.d. for a fixed threshold test).

Remark 1: Without the genie-aided threshold given to each SU selection, we may set $\boldsymbol{\eta}$ such that $\eta_i = \eta$ for all i, for some constant $\eta > 0$. Then, any selection policy which finds SU i^* can obtain a UMP (uniformly most powerful) test solution by setting η to $\sigma\Phi^{-1}(1-P_{FA})$, which maximizes the detection probability given a false alarm probability P_{FA}. Note that if $\mu_i - \mu_j > 0$, $p_i - p_j = \theta(\Phi(\frac{\mu_i-\eta}{\sigma}) - \Phi(\frac{\mu_j-\eta}{\sigma})) > 0$ in this case. However, even with $I_t = i^*$ for all $t > 1$, the UMP test does not optimize error probability.

B. UCB-FT Algorithm

We now construct the UCB-FT (Upper Confidence Bound - Fixed Threshold) algorithm which is a deterministic index-based sensing scheduling policy with a given fixed threshold

Algorithm 1 UCB-FT (UCB with Fixed Threshold)

Input: initial prior on μ, variance σ^2 and PU channel information θ
Output: scheduling sequence $\{I_t\}_{t=1}^T$

```
for t=1 to T do
  for i=1 to N do
    compute q_i^t = μ_i^t + σ_i^t Φ^{-1}(1 − 1/Kt)
  end for
  schedule SU I_t = argmax_i q_i^t
  obtain y_t and make access decision A_t according to (6)
  obtain H_t
  if H_t = 1 then
    update prior according to (10)
    increase n̂_{I_t}^t by 1
  end if
  increase n_{I_t}^t by 1
end for
```

test. In light of [26], let us first describe prior update equation for unknown parameter $\boldsymbol{\mu}$. Recall that the initial prior on $\boldsymbol{\mu}$ is Gaussian with mean $\boldsymbol{\mu_0}$ and Covariance Σ_0. For independent shadowing, we have an uncorrelated prior, $\Sigma_0 = \sigma_0^2 I$. Let us denote by $\hat{n}_i^t$ the total number of measurement samples taken from the unknown distribution the network controller has received up to time t from SU i. More formally, $\hat{n}_i^t = \sum_{j=1}^{t}\mathbb{1}(I_j = i, H_j = 1)$. Then, the uncorrelated prior is updated according to the Bayesian update rule [26], [27].

$$\mu_i^t = \frac{\xi^2\mu_i^0 + \hat{n}_i^t\bar{y}_i^t}{\xi^2 + \hat{n}_i^t}, \quad (\sigma_i^t)^2 = \frac{\sigma^2}{\xi^2 + \hat{n}_i^t} \tag{10}$$

where $\xi = \frac{\sigma}{\sigma_i^0}$, and $\bar{y}_i^t$ is the empirical mean of measurements from the unknown distribution of SU i. Since $\bar{y}_i^t \sim \mathcal{N}(\mu_i, \frac{\sigma^2}{\hat{n}_i^t})$,

$$\mu_i^t \sim \mathcal{N}\left(\frac{\xi^2\mu_i^0 + \hat{n}_i^t\mu_i}{\xi^2 + \hat{n}_i^t}, \frac{\hat{n}_i^t\sigma^2}{(\xi^2 + \hat{n}_i^t)^2}\right). \tag{11}$$

For noninformative prior, by setting $\xi \to 0$,

$$\mu_i^t \sim \mathcal{N}(\mu_i, \frac{\sigma^2}{\hat{n}_i^t}). \tag{12}$$

At each time t, the UCB-FT algorithm selects an SU with the maximum index q_i^t, i.e. $i_t = \text{argmax}_i\, q_i^t$, with

$$q_i^t = \mu_i^t + \sigma_i^t\Phi^{-1}(1 - \frac{1}{Kt}) \tag{13}$$

where K is a tunable parameter. This can be interpreted as $(1 - \frac{1}{Kt})$ upper credible limit of the unknown parameter, i.e. the true mean μ_i falls within the limit with probability $1 - \frac{1}{Kt}$. For detailed information on credible limit, see [26], [27]. On receipt of measurement y_t, UCB-FT uses fixed threshold test for access decision policy according to (6). Then, at the end of time slot t, by obtaining the true channel state H_t, the reward for time slot t is collected. Also, if the measurement sample y_t was taken from an unknown distribution, the prior should be updated according to (10). More precisely, if $I_t = i$ and $H_t = 1$, the sample mean of SU i for the next time slot should

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAES.2018.2879533, IEEE Transactions on Aerospace and Electronic Systems

be updated as $\bar{y}_i^{t+1} = \frac{\hat{n}_i^t \bar{y}_i^t + y_t}{\hat{n}_i^t + 1}$ and $\hat{n}_i^t$ should be increased by 1. For a pseudo code of the algorithm, see Algorithm 1.

C. Regret analysis of UCB-FT Algorithm

In this subsection, we analyze the performance of UCB-FT algorithm in terms of regret. As seen in (9), the regret is determined by $\mathbb{E}[n_i^T]$. We first state a theorem from [26] that will be used in the performance analysis of UCB-FT.

Theorem 1 (Bounds on the inverse Gaussian cdf [26]): The following bounds hold for the inverse cumulative distribution function of the standard Gaussian random variable for each $\alpha \in (0, \frac{1}{\sqrt{2\pi}})$ and any $\beta \geq 1.02$:

$$\Phi^{-1}(1-\alpha) < \beta \sqrt{-\log\left(-2\pi\alpha^2\right)\log(2\pi\alpha)}$$

Proof: See [26]. ∎

Now, we analyze the performance of UCB-FT algorithm. In the following theorem, an upper bound on the regret is derived and it is shown that the algorithm achieves logarithmic regret growth rate uniformly in time.

Theorem 2 (Regret of UCB-FT): For all $N > 1$, if policy UCB-FT is run on N SUs, then the expected regret after T time slots is at most

$$\inf_{\theta' \in (0,\theta)} \sum_{i=1}^{N} (p_{i^*} - p_i)\Big[\Big(\frac{2d_i}{\theta'} + \frac{1}{\sqrt{2\pi e}}\Big)\log T + \frac{d_i}{\theta'} + \frac{1}{\sqrt{2\pi e}} + \frac{2}{1 - e^{-\frac{(\theta-\theta')^2}{2}}}\Big]$$

where $d_i = \frac{4\beta^2\sigma^2}{(\mu_{i^*} - \mu_i)^2}$.

Proof: For any SU i, we upper-bound n_i^T in (8) as follows:

$$\begin{aligned} n_i^T &= \sum_{t=1}^{T} \mathbb{1}(I_t = i) \\ &\leq \sum_{t=1}^{T} \mathbb{1}(q_i^t > q_{i^*}^t) \\ &\leq l + \sum_{t=1}^{T} \mathbb{1}(q_i^t > q_{i^*}^t, n_i^{t-1} \geq l) \\ &\leq l + \sum_{t=1}^{T} \mathbb{1}(q_i^t > q_{i^*}^t, n_i^{t-1} \geq l, \hat{n}_i^{t-1} \geq 1) \\ &\quad + \sum_{t=1}^{T} \mathbb{1}(n_i^{t-1} \geq l, \hat{n}_i^{t-1} < 1) \end{aligned}$$

where l is some positive integer. At each time t, $q_i^t > q_{i^*}^t$ if at least one of the following inequalities holds.

$$\mu_{i^*}^t \leq \mu_{i^*} - C_{i^*}^t \tag{14}$$

$$\mu_i^t \geq \mu_i + C_i^t \tag{15}$$

$$\mu_{i^*} < \mu_i + 2C_i^t \tag{16}$$

where $C_i^t = \frac{\sigma}{\sqrt{\xi^2 + \hat{n}_i^t}} \Phi^{-1}(1-\alpha_t)$ with $\alpha_t = \frac{1}{Kt}$. Inequality (14) holds if

$$\begin{aligned} &\mu_{i^*}^t - \mu_{i^*} \leq -\frac{\sigma}{\sqrt{\xi^2 + \hat{n}_{i^*}^t}} \Phi^{-1}(1-\alpha_t) \\ &\Leftrightarrow z \leq -\sqrt{\frac{\hat{n}_{i^*}^t + \xi^2}{\hat{n}_{i^*}^t}} \Phi^{-1}(1-\alpha_t) + \frac{\xi^2(\mu_{i^*} - \mu_{i^*}^0)}{\sigma\sqrt{\hat{n}_{i^*}^t}} \end{aligned} \tag{17}$$

where z is a standard normal random variable. For non-informative prior, i.e. $\xi \to 0$, (17) holds if and only if $z \leq -\Phi^{-1}(1-\alpha_t)$. Therefore,

$$\mathbb{P}((14)) = \alpha_t = \frac{1}{Kt}$$

Similarly, (15) holds if

$$\begin{aligned} &\mu_i^t - \mu_i \geq \frac{\sigma}{\sqrt{\xi^2 + \hat{n}_i^t}} \Phi^{-1}(1-\alpha_t) \\ &\Leftrightarrow z \geq \sqrt{\frac{\hat{n}_i^t + \xi^2}{\hat{n}_i^t}} \Phi^{-1}(1-\alpha_t) + \frac{\xi^2(\mu_i - \mu_i^0)}{\sigma\sqrt{\hat{n}_i^t}} \end{aligned} \tag{18}$$

For noninformative prior, i.e. $\xi \to 0$, (18) holds if and only if $z \geq \Phi^{-1}(1-\alpha_t)$. Therefore,

$$\mathbb{P}((15)) = \alpha_t = \frac{1}{Kt}$$

Inequality (16) holds if

$$\mu_{i^*} - \mu_i < \frac{\sigma}{\sqrt{\xi^2 + \hat{n}_i^t}} \Phi^{-1}(1-\alpha_t) \tag{19}$$

$$\Rightarrow \frac{(\mu_{i^*} - \mu_i)^2}{4\beta^2\sigma^2}(\xi^2 + \hat{n}_i^t) < -\log(-2\pi\alpha_t^2 \log(2\pi\alpha_t^2)) \tag{20}$$

$$\Rightarrow \frac{(\mu_{i^*} - \mu_i)^2}{4\beta^2\sigma^2}(\xi^2 + \hat{n}_i^t) < 1 + 2\log T - \log 2 - \log\log T \tag{21}$$

$$\Leftrightarrow \hat{n}_i^t < \frac{4\beta^2\sigma^2}{(\mu_{i^*} - \mu_i)^2}(1 + 2\log T - \log 2 - \log\log T) - \xi^2$$

(19) ⇒ (20) follows from the inverse Gaussian tail bound shown in Theorem 1. (21) follows by setting α_t to $\frac{1}{\sqrt{2\pi e t}}$. For noninformative prior,

$$\hat{n}_i^t < d_i(1 + 2\log T - \log 2 - \log\log T)$$

where $d_i = \frac{4\beta^2\sigma^2}{(\mu_{i^*} - \mu_i)^2}$. Therefore,

$$\mathbb{P}((16)) = \mathbb{P}(\hat{n}_i^t < d_i(1 + 2\log T - \log 2 - \log\log T)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAES.2018.2879533, IEEE Transactions on Aerospace and Electronic Systems

Now, taking the expectation on n_i^T and its upper bound,

$$\begin{aligned}
\mathbb{E}[n_i^T] &\le l + \sum_{t=1}^{T} \mathbb{P}(q_i^t > q_{i^*}^t, n_i^{t-1} \ge l) \\
&\le l + \sum_{t=1}^{T} \mathbb{P}((14), n_i^{t-1} \ge l, \hat{n}_i^{t-1} \ge 1) && (22) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}((15), n_i^{t-1} \ge l, \hat{n}_i^{t-1} \ge 1) && (23) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}((16), n_i^{t-1} \ge l, \hat{n}_i^{t-1} \ge 1) && (24) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}(n_i^{t-1} \ge l, \hat{n}_i^{t-1} < 1) && (25) \\
&\le l + \frac{2}{\sqrt{2\pi e}} \sum_{t=1}^{T} \frac{1}{t} && (26) \\
&\quad + \sum_{t=1}^{T} \sum_{j=l}^{t-1} \mathbb{P}(\hat{n}_i^t \le d_i(1 + 2\log T), n_i^t = j) && (27)
\end{aligned}$$

(22) and (23) are combined and bounded by (26). (24) and (25) are combined and bounded by (27).

Upper bound of (26): From integral test,

$$(26) \le l + \frac{1}{\sqrt{2\pi e}} \int_1^T \frac{1}{t} dt = l + \frac{1}{\sqrt{2\pi e}}(1 + \log T)$$

Upper bound of (27): First, we show that $\hat{n}_i^t - \theta n_i^t$ is a martingale.

$$\begin{aligned}
\hat{n}_i^t - \theta n_i^t &= \sum_{s=1}^{t} \mathbb{1}(I_s = i)[\sum_{j=1}^{s} H_j - \sum_{j=1}^{s-1} H_j] - \theta \sum_{s=1}^{t} \mathbb{1}(I_s = i) \\
&= \sum_{s=1}^{t} \mathbb{1}(I_s = i)[\sum_{j=1}^{s} (H_j - \theta) - \sum_{j=1}^{s-1} (H_j - \theta)]
\end{aligned}$$

Let $M_s := \sum_{j=1}^{s} (H_j - \theta)$. Clearly, M_s is a martingale. Then, since $\mathbb{1}(I_s = i) \in \mathcal{F}_{s-1}$, $\hat{n}_i^t - \theta n_i^t$ is also a martingale [28].

By setting $l = \frac{d_i}{\theta'}(1 + 2logT)$ for some $\theta' < \theta$, the following holds for any $t \ge l$.

$$\begin{aligned}
&\mathbb{P}(\hat{n}_i^t \le d_i(1 + 2logT), n_i^t \ge \frac{d_i}{\theta'}(1 + 2logT)) \\
&\le \mathbb{P}(\hat{n}_i^t - \theta n_i^t \le \frac{\theta' - \theta}{\theta'} d_i(1 + 2logT)) \le \mathbb{P}(\hat{n}_i^t - \theta n_i^t \le (\theta' - \theta)t)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}(\hat{n}_i^t \le d_i(1 + logT), n_i^t \ge l) \\
&\le \sum_{t=1}^{T} \mathbb{P}(\hat{n}_i^t - \theta n_i^t \le (\theta' - \theta)t) \le \sum_{t=1}^{T} 2\exp(-\frac{(\theta' - \theta)^2 t}{2}) \\
&\le \frac{2}{1 - e^{-\frac{(\theta' - \theta)^2}{2}}}
\end{aligned}$$

The second inequality follows from Azuma-Hoeffding inequality [29].

Combining the results, we obtain an upper bound on the number of selections of SU $i \ne i^*$

$$\begin{aligned}
\mathbb{E}[n_i^t] \le \inf_{\theta' \in (0,\theta)} \sum_{i=1}^{N} [(\frac{2d_i}{\theta'} + \frac{1}{\sqrt{2\pi e}}) \log T + \frac{d_i}{\theta'} \\
+ \frac{1}{\sqrt{2\pi e}} + \frac{2}{1 - e^{-\frac{(\theta - \theta')^2}{2}}}]
\end{aligned}$$

where $d_i = \frac{4\beta^2\sigma^2}{(\mu_{i^*} - \mu_i)^2}$. This completes the proof by (9). ■

Remark 2 (The number of sub-optimal selections and θ): The leading constant of $\log T$ in the upper bound of $\mathbb{E}[n_i^t]$ is inversely proportional to $\frac{(\mu_{i^*} - \mu_i)}{\sigma}$ and monotonically decreases as θ becomes close to 1. This coincides with our intuition; Firstly, the result implies that it is easier to distinguish the best SU when there is large channel quality difference between SUs. Note that the upper bound we found is in the same order of T as the following lower bound from [9].

$$\mathbb{E}[n_i^t] \ge (\frac{1}{KL(\nu_i || \nu_{i^*})} logT + o(1)), \qquad (28)$$

where $KL(\nu_i || \nu_j)$ is the Kullback-Leibler divergence from distribution ν_i and ν_j. When ν_i are Gaussian with different mean μ_i with the same variance σ, $KL(\nu_i || \nu_{i^*}) = \frac{(\mu_{i^*} - \mu_i)^2}{2\sigma^2}$, which also appears in our upper bound. However, since sampling from the unknown distribution is limited to the time instances of busy period of PU, we can see that it also includes a parameter θ' related to PU activity. Note that we obtain lower upper bound with higher probability of busy channel. It is easy to see that the lower bound (28) also becomes lower for higher θ because even through the KL divergence between two mixed Gaussian does not have a closed form [30], it is clear the one between two mixed Gaussian ν_i and ν_{i^*} in the form of (3), i.e. $\nu_i \sim \theta\mathcal{N}(y_i|\mu_i, \sigma^2) + (1 - \theta)\mathcal{N}(y_i|0, \sigma^2)$, becomes bigger as θ increases. (However, for θ close enough to 1 or 0 such that $\theta^2 + (1 - \theta)^2 \ge p_{i^*}$, we can simply use randomized access policy $A_t = 1$, w.p. θ for all $t \ge 1$, without sensing). For a fixed θ, θ' balances between the constant term and the log term. The leading constant of $\log T$ becomes smaller as θ' approaches to θ, at the cost of larger constant term.

Remark 3 (Comparison with 0-1 feedback): In UCB-FT, it was assumed that the selected SU at each time t reports its measurement result y_t and the index q_i^t associated with each SU i is expressed as a function of $\bar{y}_i^t$. However, it is also possible for the selected SU to make a local decision based on its measurement and report its 0-1 binary result to the network controller. This is more desirable when the feedback resource is limited or the noise model is only known at the SU side. By selecting SU i at time t, the network controller observes a reward x_i^t drawn from an unknown Bernoulli distribution $\mathcal{B}(p_i)$ instead of y_i^t from an unknown Normal distribution $\mathcal{N}(\mu_i, \sigma^2)$. Since Bernoulli reward has a support [0,1], the network controller can use an allocation algorithm such as UCB-1 from [11] for sensor selection by setting the index q_i^t as follows:

$$q_i^t = \bar{x}_i^t + \sqrt{\frac{2\log t}{n_i^t}}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAES.2018.2879533, IEEE Transactions on Aerospace and Electronic Systems

where $\bar{x}_i^t$ is the average reward obtained from SU i. Note that the index used for UCB-FT has a similar form to this with $\bar{y}_i^t$, $\hat{n}_i^t$ and $\sigma^2\Phi^{-1}(1-\frac{1}{Kt})$ in place of $\bar{x}_i^t$, n_i^t and $\log t$. Using UCB-1, the expected number of selections of sub-optimal SU i is upper-bounded by:

$$\mathbb{E}[n_i^T] \leq \frac{8}{(p_{i^*}-p_i)^2}\log T + 1 + \frac{\pi^2}{3}.$$

Note that the leading constant of $\log T$ is inversely proportional to $(p_{i^*}-p_i)^2$, which is always smaller than $(\mu_{i^*}-\mu_i)^2$ for a given $\boldsymbol{\mu}$ from the definition of p_i in (8). Therefore, by learning $\boldsymbol{\mu}$ instead of $\mathbf{p}$, we can expect to distinguish sub-optimal arms more easily, and as the channel quality improves, the effect will become more substantial. However, it must also be noted that in a high SNR environment in which μ is relatively large compared to σ, the difference between expected rewards from optimal and sub-optimal selections can become negligible as all SUs have very low error probability.

IV. Sequential sensor selection with a sample mean based threshold test

A. Definition of Regret

In the previous section, it is assumed that the access policy is given so that we can focus more on the sensing scheduling. However, to achieve an optimal solution of the original problem, optimal access policy needs to be employed along with optimal sensor selection. For any given selection, it is known that LLR (Log Likelihood Ratio) test minimizes the probability of error. However, the threshold of LLR test is a function of the unknown parameter μ. More specifically, if SU i is scheduled for sensing at time slot t, LLR test will give the following threshold test with unknown μ_i. Note that $\eta_i^* = \text{argmin}_\eta\, 1 - p_i(\eta) = \text{argmin}_\eta\, 1 - ((1-\theta)\Phi(\frac{\eta}{\sigma}) + \theta\Phi(\frac{\mu_i-\eta}{\sigma}))$.

$$y_i^t \underset{0}{\overset{1}{\gtrless}} \eta_i^* = \frac{\mu_i}{2} + \log\frac{1-\theta}{\theta} \tag{29}$$

Now, we define the optimal reward with optimal threshold test with known $\boldsymbol{\mu}$. Let i^* denote the index of SU with largest μ, i.e. $i^* = \text{argmax}_i\, \mu_i$. Then, the optimal policies are always picking SU i^* for sensing and making access decision using threshold test with $\eta^* = \frac{\mu_{i^*}}{2} + \log\frac{1-\theta}{\theta}$. From (8),

$$R_{\text{opt}}^T = p^*T = [(1-\theta)\Phi(\frac{\mu_{i^*}+2a}{2\sigma}) + \theta\Phi(\frac{\mu_{i^*}-2a}{2\sigma})]T \tag{30}$$

where $a = \log\frac{1-\theta}{\theta}$.

Obviously, if we use a stationary threshold test with some $\boldsymbol{\eta} \neq \boldsymbol{\eta}^*$, the optimal reward p^*T is not equal to $p_{i^*}T$. Therefore, it is possible that the regret increases linearly in time horizon T because $(p^* - p_i) > 0$ for all i ***including*** i^*. Note that the number of selections of SU i^* increases linearly (i.e. $\mathbb{E}[n_{i^*}^T] = O(T)$) using UCB-FT. The logarithmic regret shown in Theorem 2 is due to the assumption that $(p^* - p_{i^*}) = 0$, which is not feasible in practice.

Algorithm 2 UCB-LLR (UCB with LLR test)

6: obtain y_t and make access decision A_t as follows.

$$y_t \underset{A_t=0}{\overset{A_t=1}{\gtrless}} \frac{\mu_{i_t}^t}{2} + \log\frac{1-\theta}{\theta}$$

B. UCB-LLR Algorithm

From (10), we know μ_i^t (the estimate of μ_i at time t) will converge to true μ_i as $\hat{n}_i^t \to \infty$. Therefore, in algorithm 2, we slightly modify the UCB-FT algorithm so that instead of using fixed threshold for access policy, we use the access policy (29) with μ_i replaced by μ_i^t. The sensing scheduling policy remains the same, which means the selection sequence is equivalent to that of UCB-FT. The only difference would be the reward drawn from each selection. We recall that for noninformative prior, μ_i^t is equivalent to the sample mean of observations on the active PU channel taken from SU i upto time t.

C. Regret Analysis of UCB-LLR Algorithm

A policy with a sub-linear growth rate of regret is long-run average optimal in the sense that it achieves the same maximum average expected reward as in the known parameter case asymptotically [14]. However, the slower the regret growth is, the faster the convergence to this maximum average reward, indicating a more effective learning ability of the policy. In the following theorem, we show that the growth of the regret of UCB-LLR is sub-linear in time horizon T, $O(T^{\frac{2}{3}+\epsilon})$ for some $\epsilon > 0$ to be more specific.

Theorem 3 (Regret of UCB-LLR): For all $N > 1$, if policy UCB-LLR is run on N SUs, then the expected regret after T time slots is at most

$$\inf_{m,\theta' \in (0,1)} \frac{c_1}{\theta'\sqrt{\theta'}} T^{1-\frac{m}{2}}(\sqrt{logT}+1) + \frac{1}{\theta'}T^m + \sum_{i=1}^{N} c_2 logT + \sum_{i\neq i^*}(\frac{2d_i}{\theta'}\log T + \frac{d_i}{\theta'}) + \frac{2N}{\sqrt{2\pi e}} + \frac{2N}{1-e^{-\frac{(\theta-\theta')^2}{2}}}$$

where $c_1 = \beta\sigma$ and $c_2 = \frac{2}{\sqrt{2\pi e}}$ and $d_i = \frac{4\beta^2\sigma^2}{(\mu_{i^*}-\mu_i)^2}$.

Proof: Firstly, we note that the equalities in (9) do not hold as the reward process is no longer i.i.d., thus $\mathbb{E}[X_i^t|\mathcal{F}_{t-1}] \neq \mathbb{E}[X_i^t]$. However, even though the Bernoulli reward process X_t is not i.i.d., we can still benefit from the fact that the sensing samples are taken from a stationary distribution of (1). Observe that $\mathbb{E}[X_i^t|\mathcal{F}_{t-1}] = \mathbb{E}[\mathbb{1}(y_i^t \geq \frac{\mu_i^t}{2} + \log\frac{1-\theta}{\theta}, H_t = 1) + \mathbb{1}(y_i^t < \frac{\mu_i^t}{2} + \log\frac{1-\theta}{\theta}, H_t = 0)|\mathcal{F}_{t-1}]$. Since $\mu_i^t \in \mathcal{F}_{t-1}$ and H_t and y_i^t are independent of $\mathcal{F}_{t-1}$, we obtain the following equality.

$$R_{\text{opt}}^T - R^T = p^*T - \sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}[\mathbb{E}[X_t\mathbb{1}(I_t=i)|\mathcal{F}_{t-1}]] \tag{31}$$

$$\leq \sum_{t=1}^{T}\mathbb{E}[(p(\mu_{i^*}) - p(\mu_{i^*}^t))\mathbb{1}(I_t=i^*)] + \sum_{t=1}^{T}\sum_{i\neq i^*}\mathbb{E}[(p(\mu_{i^*}) - p(\mu_i^t))\mathbb{1}(I_t=i)] \tag{32}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAES.2018.2879533, IEEE Transactions on Aerospace and Electronic Systems

where $p(\mu_{i^*}) = p^*$ and $p(\mu_i^t) = (1-\theta)\Phi(\frac{\mu_i^t+2a}{2\sigma}) + \theta\Phi(\frac{2\mu_i-2a-\mu_i^t}{2\sigma})$ from (9) and (31). To emphasize the dependency of p_i on μ_i^t, we use $p(\mu_i^t)$ instead of p_i here. As the regret for $i \neq i^*$ is upper bounded by $\mathbb{E}[n_i^T]$ which has been proved to grow in the order of $\log T$ by Theorem 2, we focus on bounding the first addend of (32).

$$\begin{aligned}
&\sum_{t=1}^{T}(p(\mu_{i^*}) - p(\mu_{i^*}^t))\mathbb{1}(I_t = i^*) \\
&\leq l + \sum_{t=1}^{T}(p(\mu_{i^*}) - p(\mu_{i^*}^t))\mathbb{1}(I_t = i^*, n_{i^*}^{t-1} \geq l) \qquad (33)\\
&\leq l + \sum_{t=1}^{T}(p(\mu_{i^*}) - p(\mu_{i^*}^t))\{\mathbb{1}(n_{i^*}^{t-1} \geq l, \hat{n}_{i^*}^{t-1} \geq \theta' l) \\
&\quad + \mathbb{1}(n_{i^*}^{t-1} \geq l, \hat{n}_{i^*}^{t-1} < \theta' l)\} \\
&\leq l + \sum_{t=1}^{T}|\mu_{i^*} - \mu_{i^*}^t|\mathbb{1}(\hat{n}_{i^*}^{t-1} \geq \theta' l, |\mu_{i^*} - \mu_{i^*}^t| \leq C_{i^*}^t) \\
&\quad + \sum_{t=1}^{T}\mathbb{1}(\hat{n}_{i^*}^{t-1} \geq \theta' l, |\mu_{i^*} - \mu_{i^*}^t| \geq C_{i^*}^t) \\
&\quad + \sum_{t=1}^{T}\mathbb{1}(n_{i^*}^{t-1} \geq l, \hat{n}_{i^*}^{t-1} < \theta' l) \qquad (34)
\end{aligned}$$

The inequality (33) holds because

$$\begin{aligned}
&\sum_{t=1}^{T}(p(\mu_{i^*}) - p(\mu_{i^*}^t))\mathbb{1}(I_t = i^*, n_{i^*}^{t-1} < l) \\
&\leq \sum_{t=1}^{T}\mathbb{1}(I_t = i^*, n_{i^*}^{t-1} < l) \leq l.
\end{aligned}$$

The inequality (34) holds because $|p(\mu_{i^*}) - p(\mu_{i^*}^t)| \leq \frac{1}{\sqrt{2\pi}}|\mu_{i^*} - \mu_{i^*}^t|$ and $|p(\mu_{i^*}) - p(\mu_{i^*}^t)| \leq 1$.

Taking expectation on both sides,

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}[(p(\mu_{i^*}) - p(\mu_{i^*}^t))\mathbb{1}(I_t = i^*)] \leq l \\
&\quad + \sum_{t=1}^{T}\sqrt{\frac{\sigma^2}{\theta' l}(\Phi^{-1}(1-\frac{1}{Kt}))^2}\mathbb{P}(\hat{n}_{i^*}^{t-1} \geq \theta' l, |\mu_{i^*} - \mu_{i^*}^t| \leq C_{i^*}^t) \qquad (35)\\
&\quad + \sum_{t=1}^{T}\mathbb{P}(z \leq \Phi^{-1}(1-\frac{1}{Kt})) \qquad (36)\\
&\quad + \sum_{t=1}^{T}\mathbb{P}(n_{i^*}^{t-1} \geq l, \hat{n}_{i^*}^{t-1} < \theta' l) \qquad (37)
\end{aligned}$$

Upper bound of (35): By setting $l = \frac{T^m}{\theta'}$ $(0 < m < 1)$ and $K = \sqrt{2\pi e}$, (35) is bounded by

$$\begin{aligned}
&\sum_{t=1}^{T}\sqrt{\frac{\sigma^2}{\theta' l}(\Phi^{-1}(1-\frac{1}{Kt}))^2} \\
&\leq \frac{c_1}{\sqrt{T^m}}\sum_{t=1}^{T}(1+\sqrt{2logt}) \leq c_1 T^{1-\frac{m}{2}}(1+\sqrt{logT})
\end{aligned}$$

where $c_1 = \beta\sigma\theta'^{-\frac{3}{2}}$. The first inequality follows from Theorem 1.

Upper bound of (36):

$$(36) = \sum_{t=1}^{T}\frac{1}{Kt} \leq \frac{2}{\sqrt{2\pi e}}(1+logT).$$

Upper bound of (37): Similar to the upper bound of (27), for any $t \geq l$.

$$\begin{aligned}
(37) &\leq \sum_{t=l}^{T}\mathbb{P}(\hat{n}_i^t - \theta n_i^t \leq \frac{\theta' - \theta}{\theta'}T^m) \\
&\leq \sum_{t=1}^{T}\mathbb{P}(\hat{n}_i^t - \theta n_i^t \leq (\theta' - \theta)t) \leq \frac{2}{1 - e^{-\frac{(\theta'-\theta)^2}{2}}}
\end{aligned}$$

Combining the results, the regret of SU i^* is bounded by

$$\begin{aligned}
&\frac{c_1}{\theta'\sqrt{\theta'}}T^{1-\frac{m}{2}}(\sqrt{logT}+1) + \frac{1}{\theta'}T^m + c_2 logT \\
&\qquad + 2(\frac{1}{1 - e^{-\frac{(\theta'-\theta)^2}{2}}} + \frac{1}{\sqrt{2\pi e}}) \qquad (38)
\end{aligned}$$

where $c_1 = \beta\sigma$ and $c_2 = \frac{2}{\sqrt{2\pi e}}$. Combining (38) and the bound from Theorem 2 for $i \neq i^*$ completes the proof. ■

Remark 4: Since $\forall\epsilon \in \mathbb{R}^+$, $logT < \frac{T^\epsilon}{\epsilon}$, (38) can be written in the following form.

$$\frac{1}{\theta'}T^m + \frac{c_1}{\theta'\sqrt{\epsilon\theta'}}(T^{1-\frac{m}{2}+\frac{\epsilon}{2}} + T^{1-\frac{m}{2}}) + c_2 logT + c_3 \qquad (39)$$

where c_3 is the constant term of (38). (39) has the lowest order in T when $m = \frac{2}{3} + \frac{\epsilon}{3}$. Then, the bound becomes

$$c_4 T^{\frac{2}{3}+\frac{\epsilon}{3}} + c_5 T^{\frac{1}{3}-\frac{\epsilon}{3}} + c_2 logT + c_3 \qquad (40)$$

where $c_4 = \frac{1}{\theta'} + \frac{c_1}{\theta'\sqrt{\epsilon\theta'}}$ and $c_5 = \frac{c_1}{\theta'\sqrt{\epsilon\theta'}}$. Note that c_4 and c_5 is inversely proportional to ϵ. Therefore, the upper bound on the regret has lower order in T as ϵ approaches 0, at the cost of larger leading constant.

Remark 5 (Multiple SU selection): In this paper, we consider a single SU selection at a time. If we select M SUs, $M \geq 2$, the problem will be in the form of combinatorial multi-armed bandit [16]. By defining a set of arms as a super arm, it is possible to treat each super arm as a classical arm and apply the same algorithms, i.e. choosing the M SUs with the M highest q_i^t at each time t. This approach will result in an exponential number of arms. However, considering that we are more interested in the case where only a few SUs can be scheduled at a time, this may be tolerable. If $M = N$, we no longer need an selection algorithm, and the problem reduces to finding an optimal threshold for the access decision. Since in LDA, the optimal threshold test is equivalent to comparing $\mu^T y$ with its threshold, learning μ will be equivalent to finding an optimal weight vector for the data fusion [7]. However, as M becomes large, the difference between the optimal solution and the one using approximate threshold test can become negligible.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAES.2018.2879533, IEEE Transactions on Aerospace and Electronic Systems

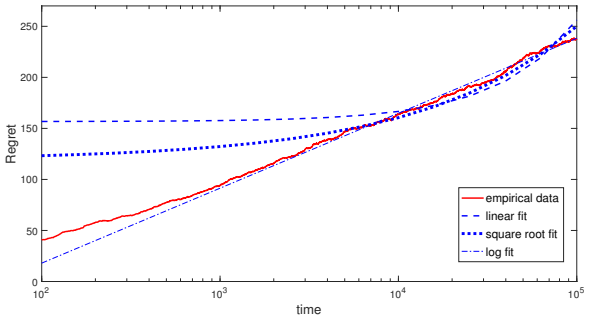


Fig. 5: Best fit to the empirical data from UCB-LLR

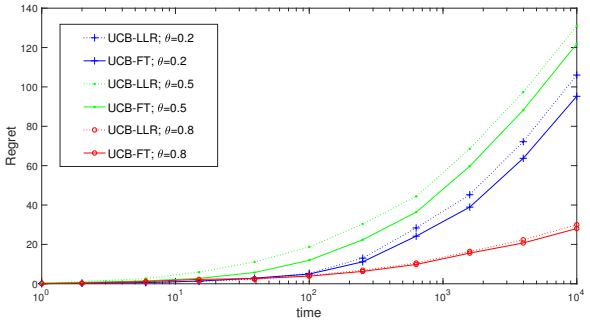


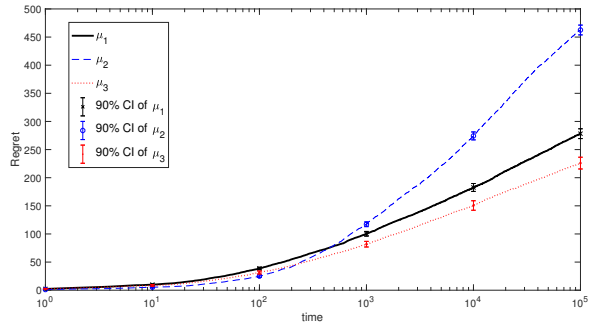
Fig. 6: The effect of the probability of channel availability θ

V. Performance Evaluation

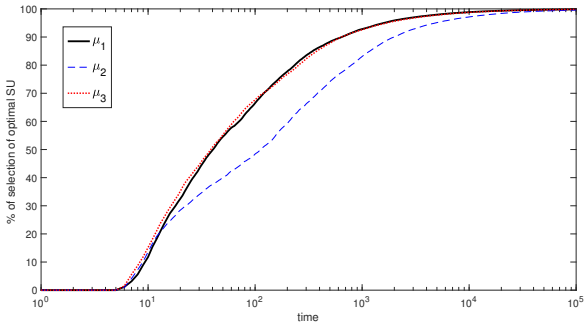
In this section, the algorithms of Section III and Section IV are evaluated numerically. We consider an SU network consisting of 12 SUs distributed randomly over 10km distance from PU transmitter. Noise variance $\sigma^2 = 1.0$ and μ is a realization of random shadow fading with fixed dB-spread $\sigma_0^2 = 2.0$. Except for the result in Fig. 6, $\theta = 0.5$. For each experiment, we track the accumulated regret and the percentage of selections of optimal SU. The plots show these quantities on semi-log scale for 10^5 time slots averaged over 500 different runs.

The result in Fig. 5 shows the empirical behavior of UCB-LLR. We fit three regret models to the data: (1) $aT + b$, (2) $a\sqrt{T}+b$ and (3) $a \log T + b$. The best fit with minimum squared error is depicted in Fig. 5. From which, we can see that UCB-LLR has sub-linear regret in T. In fact, the results resemble logarithmic growth very closely.

The result in Fig. 6 shows the effect of the probability of channel availability on the algorithm performance. The regret achieved under both UCB-FT and UCB-LLR is presented for different θ's. Earlier in Remark 2, we discussed that higher θ provides more samples from the unknown distribution of $y_t|H_t = 1$. So, from the aspect of learning, the expected number of selections of sub-optimal SUs decreases as θ approaches to 1. However, for the regret, we need a more careful examination as the expected reward p_i drawn from each selection of SU i is also a function of θ. From the definition of p_i in (8), we know that $p_i - p_{i^*}$ is a symmetric bell shaped curve in θ, centered at $\theta = \frac{1}{2}$. This means that due to small regret per selection, the accumulated regret for some $\theta < \frac{1}{2}$ can be lower than the one for $\theta = \frac{1}{2}$ even with more selections. The result for $\theta = 0.2$ being better than $\theta = 0.5$ can be explained by this. Clearly, for large θ, not only have we better learning from more sensing samples from $y_t|H_t = 1$ but the regret from each selection is also small i.e. at the tail of the curve (8) in θ. The result in Fig. 6 also provides a comparison of empirical behaviors of UCB-FT and UCB-LLR. Even though Theorem 3 only proved the sub-linear growth of regret in T of UCB-LLR, Fig. 6 shows that it is very close to that of UCB-FT, which has a logarithmic regret growth rate in T.



(a) Regret



(b) % of selection of optimal SU

Fig. 7: The effect of scaling and transition of $\boldsymbol{\mu}$ on the algorithm performance

In Fig. 7, we observe the effect of different settings of unknown parameter $\boldsymbol{\mu}$ on the regret performance with 90% confidence intervals marked at $t = \{10, 10^2, 10^3, 10^4, 10^5\}$. We consider three different SU networks having the mean vector $\boldsymbol{\mu^1}$ (network 1), $\boldsymbol{\mu^2}$ (network 2) and $\boldsymbol{\mu^3}$ (network 3) for the distribution of $\mathbf{y}_t|H_t = 1$. We choose a network 1 with a random $\boldsymbol{\mu^1}$ as a reference and generate $\boldsymbol{\mu^2}$ such that $\mu^2_{i^*} = \mu^1_{i^*}$ for $i^* = \text{argmax}\, \mu^1_i$ and $\mu^2_i - \mu^2_{i^*} = \frac{1}{2}(\mu^1_i - \mu^1_{i^*})$ for all $i \neq i^*$. Note that this setting can happen when the characteristics of an SU network such as network coverage or shadowing variance change. As discussed in Remark 3, since the KL divergence between the distribution of $y^t_i|H_t = 1$ and $y^t_{i^*}|H_t = 1$ of the network 2 has been reduced from that of network 1, we have a larger leading constant of $logT$ in the regret bound and in the number of selections of sub-optimal SUs. The result of $\boldsymbol{\mu^2}$ in Fig. 7 coincides with this observation. It is shown in Fig. 7 that we have a higher regret growth rate and a lower convergence rate to the optimal SU selection with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAES.2018.2879533, IEEE Transactions on Aerospace and Electronic Systems

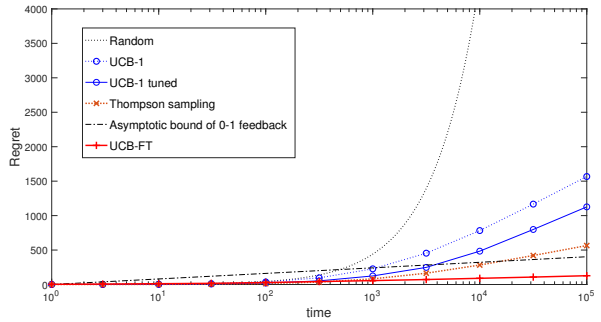


Fig. 8: Performance comparison between UCB-FT and 0-1 feedback algorithms

$\boldsymbol{\mu^2}$. However, as discussed in Fig. 6, the KL divergence of $y_i^t|H_t = 1$ does not tell everything about the regret. To show this, we take $\boldsymbol{\mu^3}$ such that $\mu_i^3 = \mu_i^1 + 0.5$, for all i. Note that this can be interpreted as the channel gain being improved by 0.5 for all SUs. Under this setting, the KL divergence between the distribution of $y_i^t|H_t = 1$ and $y_{i^*}^t|H_t = 1$ of the SUs in the network 3 remains the same as network 1. However, similar to the effect of θ on the regret for a given $\mu_{i^*} - \mu_i$, for a given θ, any change in μ_i including transitions affect the value of $p_{i^*} - p_i$ due to the non-linear relationship between μ_i and p_i. Therefore, even a simple shifting of $\boldsymbol{\mu^1}$ to $\boldsymbol{\mu^3}$ can cause changes in the regret growth rate. The results for $\boldsymbol{\mu^3}$ in Fig. 7 explains this. The number of selections of the optimal SU is almost the same for $\boldsymbol{\mu^1}$ and $\boldsymbol{\mu^3}$ whereas $\boldsymbol{\mu^3}$ has smaller regret than $\boldsymbol{\mu^1}$.

In Fig. 8, we compare the empirical behavior of UCB-FT and some existing selection algorithms UCB-1 [11], UCB-1 tuned and Thompson sampling [31]. Note that UCB-1, UCB-1 tuned, and Thompson sampling can be considered as a 0-1 feedback while UCB-FT sends back the measurement data y_t to the network controller. Using 0-1 feedback, if SU i is selected for sensing at time slot t, it measures the PU signal, make a local binary decision and reports the result to the network controller. The result in Fig. 8 shows how we can benefit from an underlying structure of reward process, taking advantage of relatively large KL divergence between two Gaussian distributions of $y_i^t|H_t = 1$ and $y_{i^*}^t|H_t = 1$ compared to the two Bernoulli distributions x_i^t and $x_{i^*}^t$, where $x_t = \mathbb{1}(y_i^t \geq \eta_i, H_t = 1) + \mathbb{1}(y_i^t < \eta_i, H_t = 0)$. By using UCB-FT, it is shown that a regret growth rate which is even lower than the theoretically proven lower bound of binary feedback [9], i.e. $\sum_{i \neq i^*} \frac{p_{i^*} - p_i}{KL(p_i||p_{i^*})} logT$, is achievable asymptotically as we send back y_t instead of binary decisions. However, it should be noted that in practice, a 0-1 feedback also needs to employ an access policy that can operate under uncertainty such as the one proposed in our Algorithm 2.

VI. Conclusion

In this work, we develop a sequential sensor selection and channel access algorithm without a priori information on the underlying measurement distribution from which the sensing samples are taken, with the objective of maximizing total number of correct decisions on channel availability over finite time period. To overcome the uncertainty issue, we formulate our problem in the context of MAB problems and propose a heuristic based access decision rule and a sensor selection algorithm. We first show that the number of sub-optimal selections of our proposed algorithm grows at most logarithmically in time period T. Then, combined with our access decision rule, we finally show that the algorithm achieves sub-linear accumulated regret in time period T, i.e. long-run average optimal. The performance of our algorithm is evaluated through simulations under different channel settings.

Our study could be extended in several ways. Since channel correlation can be easily incorporated in the prior distribution of $\boldsymbol{\mu}$ in our signal model, further study could elaborate on this point providing a comprehensive design for cooperative sensing. Further research could also take into account more complicated settings such as (unknown) Markov PU channel or fast user mobility. In particular, as we assumed that the shadowing effect is time-invariant during the time horizon T, the extent of allowable environmental variation and the minimal achievable regret under high user mobility will be studied in our future work.

References

[1] Hessar, F. and Roy, S., "Spectrum sharing between a surveillance radar and secondary Wi-Fi networks." IEEE Transactions on Aerospace and Electronic Systems 52.3 (2016): 1434-1448.

[2] Mahal, A., et al., "Spectral coexistence of MIMO radar and MIMO cellular system." IEEE Transactions on Aerospace and Electronic Systems 53.2 (2017): 655-668.

[3] Atapattu, S., Tellambura, C. and Jiang, H., "Performance of an energy detector over channels with both multipath fading and shadowing," *IEEE Transactions on Wireless communications*, vol. 9, no. 12, pp. 3662–3670, 2010.

[4] Min, A. and Shin, K., "An optimal sensing framework based on spatial rss-profile in cognitive radio networks," in *Sensor, Mesh and Ad Hoc Communications and Networks, 2009. SECon'09. 6th Annual IEEE Communications Society Conference on*. IEEE, 2009, pp. 1–9.

[5] Li, S., et al. "Maximizing system throughput by cooperative sensing in cognitive radio networks."IEEE/ACM Transactions on Networking (TON)22.4 (2014): 1245-1256.

[6] Unnikrishnan, J. and Veeravalli, V., "Cooperative sensing for primary detection in cognitive radio." IEEE Journal of selected topics in signal processing 2.1 (2008): 18-27.

[7] Cacciapuoti, A., Akyildiz, I. and Paura, L., "Correlation-aware user selection for cooperative spectrum sensing in cognitive radio ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 2, pp. 297–306, 2012.

[8] Poor, H., *An introduction to signal detection and estimation*. Springer Science & Business Media, 2013.

[9] Lai, T. and Robbins, H., "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[10] López-Martínez, M., Alcaraz, J., Badia, L. and Zorzi, M., "Multi-armed bandits with dependent arms for cooperative spectrum sharing," in *Communications (ICC), 2015 IEEE International Conference on*. IEEE, 2015, pp. 7677–7682.

[11] Auer, P., Cesa-Bianchi, N. and Fischer, P., "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[12] Anandkumar, A., Michael, N. and Tang, A., "Opportunistic spectrum access with multiple users: Learning under competition," in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[13] Lai, L., El Gamal, H., Jiang, H. and Poor, H., "Cognitive medium access: Exploration, exploitation, and competition," *IEEE Transactions on Mobile Computing*, vol. 10, no. 2, pp. 239–253, 2011.

[14] Liu, H., Liu, K. and Zhao, Q., "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2013.

[15] Liu, K. and Zhao, Q., "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[16] Chen, W., Wang, Y. and Yuan, Y., "Combinatorial multi-armed bandit: General framework, results and applications," in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 151–159.

[17] Lunden, J., Koivunen, V. and Poor, H., "Spectrum exploration and exploitation for cognitive radio: Recent advances," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 123–140, 2015.

[18] Ouyang, W., Eryilmaz, A. and Shroff, N., "Downlink scheduling over markovian fading channels," *IEEE/ACM Transactions on Networking*, vol. 24, no. 3, pp. 1801–1812, 2016.

[19] Ferrari, L., Zhao, Q. and Scaglione, A., "Utility Maximizing Sequential Sensing Over a Finite Horizon." IEEE Transactions on Signal Processing 65.13 (2017): 3430-3445.

[20] Sangare, F. and Han, Z., "Joint optimization of cognitive RF energy harvesting and channel access using Markovian Multi-Armed Bandit problem."Communications Workshops (ICC Workshops), 2017 IEEE International Conference on. IEEE, 2017.

[21] Zheng, R. and Hua, C., "Spectrum Sensing and Access in Cognitive Radio Networks."Sequential Learning and Decision-Making in Wireless Resource Management. Springer International Publishing, 2016. 61-69.

[22] Bagheri, S. and Scaglione, A., "The restless multi-armed bandit formulation of the cognitive compressive sensing problem."IEEE Transactions on Signal Processing63.5 (2015): 1183-1198.

[23] Norinho, D., "Dynamic Learning Gaussian Process Bandits." RN 12 (2012): 15.

[24] Chen, Y., Zhao, Q. and Swami, A., "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2053–2071, 2008.

[25] Jouini, W., Moy, C., Palicot, J., *et al.*, "Upper confidence bound algorithm for opportunistic spectrum access with sensing errors," in *6th International ICST Conference on Cognitive Radio Oriented Wireless Networks and Communications, Osaka, Japan*, vol. 17, 2011.

[26] Reverdy, P., Srivastava, V. and Leonard, N., "Modeling human decision making in generalized gaussian multiarmed bandits," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 544–571, 2014.

[27] Berger, J., *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[28] Durrett, R., *Probability: theory and examples*. Cambridge university press, 2010.

[29] Azuma, K., "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal, Second Series*, vol. 19, no. 3, pp. 357–367, 1967.

[30] Hershey, J. and Olsen, P., "Approximating the Kullback Leibler divergence between Gaussian mixture models." Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Vol. 4. IEEE, 2007.

[31] Chapelle, O. and Li, L., "An empirical evaluation of thompson sampling." Advances in neural information processing systems. 2011.

Jihyun Lee (S'16) received her BS and MS degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2007 and 2009, respectively . She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA. Her research interests include cognitive radio, network optimization, and resource management in wireless networks.

Eylem Ekici (S'99-M'02-SM'11-F'17) received his BS and MS degrees in Computer Engineering from Bogazici University, Istanbul, Turkey, in 1997 and 1998, respectively. He received his Ph.D. degree in Electrical and Computer Engineering from Georgia Institute of Technology, Atlanta, GA, in 2002. Currently, he is a professor in the Department of Electrical and Computer Engineering of The Ohio State University, Columbus, OH. He is an Associate Editor-in-Chief of IEEE Transactions on Mobile Computing, and a former associate editor of IEEE/ACM Transactions on Networking, IEEE Transactions on Mobile Computing, and Computer Networks Journal (Elsevier). He served as the general co-chair of ACM MobiCom 2012. He was also the TPC Co-Chair of IEEE INFOCOM 2017. Dr. Ekici is the recipient of the 2008 and 2014 Lumley Research Award of the College of Engineering at OSU. He also received the 2016 Harrison Faculty Award for Excellence in Engineering Education. Dr. Ekici's current research interests include cognitive radio networks, vehicular communication systems, and next generation wireless systems, with a focus on algorithm design, medium access control protocols, resource management, and analysis of network architectures and protocols. He is an IEEE Fellow and a member of ACM.